RESEARCH ARTICLE   Legume Research- An International Journal

Establishment of Detection Model of Soybean Quality Traits by Near Infrared Spectroscopy

Weiran Gao[1#], Ronghan Ma[1#], Aohua Jiang[1], Jiaqi Liu[1], Pingting Tan[1], Fang Liu[1], Jian Zhang[1]

10.18805/LRF-760

ABSTRACT

Background: Rapid prediction with near infrared (NIR) spectroscopy on quality traits is pretty popular recently, for the convenience and simple operation. But to make good use of this technology, precise and suitable calibration equations are very important to get dependable result. In this study, we mostly refer to the building of the equation and how the pretreatment effect them.

Methods: In this paper, near infrared (NIR) spectroscopy was used to simultaneously predict the quality traits of soybean, including oil content, protein content, oleic acid content, linoleic acid content, stearic acid content. Near infrared spectral data of a total of 112 samples is collected from given materials in Chongqing. Samples were scanned from 1000 nm to 2500 nm using a monochromator instrument (SuperNIR-2700). Calibration equations were developed from NIR data using partial least squares (PLS) regression with internal cross validation. In addition, in this study, we also cover the affection of different pre-treatments to the different calibration equations predicting different quality traits. And measure the effect with three indicators including R, SECV and RPD.

Result: Eventually we find the most suitable combination of pre-treatments for each calibration equation predicting a certain trait soybean. The present study would lay the foundations of rapid detection of quality traits in soybean.

Key words: Near infrared spectroscopy, Pre-treatment, Quality traits, Soybean.

INTRODUCTION

Improvement on quality traits has always been one of the most significant targets in crops breeding. In soybean [*Glycine max* (L.) Merr], oil content (Clemente and Cahoon 2009), protein content (Medic and Atkinson, 2014) and oil composition are especially important. Soybean oil is mainly composed of five fatty acids, palmitic, stearic, oleic, linoleic and linolenic acids (Kinney and Knowlton, 1998), Oleic acid is a monounsaturated fatty acid that facilitates oxidative stability for increased shelf life, heat stability in cooking and health benefits (Zambelli, 2020). Thus, the improvement on those quality traits are surely desirable.

At present, classic chemistry methods is necessary to measure the quality traits of crops. Like the Kjeldahl method for protein content and the Soxhlet extraction method for oil content (Jung and Rickert, 2003). Although these methods have the advantages of low cost, simple operation and high accuracy, they are time-consuming and labor-intensive and will cause environmental pollution. Meanwhile the method of near infrared spectroscopy takes little time and thus become more and more popular and widely used since it was developed (Corte and Blasco, 2019). Nowadays, the application of near-infrared spectroscopy technology to analyze and measure various quality traits of crops is common. It can achieve fully automatic operation, with relatively small human error caused during the entire measurement process and has high precision and good reproducibility. It is not only widely used in crop quality prediction, but also widely applicable in various biological, medical and other fields (McClure, 2003, Nicolaï and Beullens, 2007). This technology only requires a large amount of basic work such as collection, calibration and equation establishment to be completed in the preliminary work. After the model is established, unknown samples can be measured using the established model and collected spectra. The data of the required chemical components of the sample can be obtained within 1 minute without causing damage or pollution to the test samples.

Near infrared spectroscopy (NIR), the light in the wavelength range of 780 nm-2526 nm contains information of the relative proportions of C-H, N-H and O-H bonds which are the primary structural components of organic molecules (Nikolić and Jović, 2007). By analyzing the correlation between sample composition, as determined by defined reference chemical methods and the absorption of light at different wavelengths in the near infrared region measured in the certain environment by the certain machine, we could build a calibration equation for this once for all.

[1]College of Agronomy and Biotechnology, Southwest University, Chongqing 400715, China.
[#]These authors contributed equally to this work.

Corresponding Author: Jian Zhang, College of Agronomy and Biotechnology, Southwest University, Chongqing 400715, China.
Email: zhangjianswau@126.com

How to cite this article: Gao, W., Ma, R., Jiang, A., Liu, J., Tan, P., Liu, F. and Zhang, J. (2024). Establishment of Detection Model of Soybean Quality Traits by Near Infrared Spectroscopy. Legume Research. doi: 10.18805/LRF-760.

Submitted: 20-07-2023 **Accepted:** 25-09-2023 **Online:** 04-01-2024

However, NIR method is not perfect, first we could not guarantee the consistence of every scan. There are a lot of factors effecting the spectrum (Qiao and Mu, 2021). Under different environmental conditions, the near-infrared spectrum of soybeans will have some slight differences. As a result, to insure the accuracy, building a particular equation to suit the environment is necessary.

MATERIALS AND METHODS

Sample

The representativeness of the sample determines the effectiveness of modeling. In order to obtain representative samples, 112 soybean materials with significant differences in quality traits were selected for this experiment in the laboratory; Randomly select 10 samples for the correction models of protein and oil content as the validation set. The samples were planted in summer of 2022 in Chongqing and sown in single row, with row length of 1m, row width of 0.5 m, plant spacing of 0.2 m. All samples were conducted with general field management.

Instrument

RT-01A 50G Western Medicine Crusher for grinding the materials. YG-2 cable extractor, K1100 fully automatic Kjeldahl nitrogen analyzer and Agilent Technologies GC (Gas chromatography) system for the determination of quality traits; SupNIR-2700 system (Including detector and software) for the scanning and analysis of spectrum.

Near infrared scanning

Samples were scanned with SupNIR-2700 system in the range of 1000-1799 nm. Each sample is repeatedly scanned 3 times. A reference scan was taken once in every 30 sample scans. Samples were temperature equilibrated at 26°C in the instrument before scanning. Spectral data were collected using SuperNIR-2700 software.

Chemometrics and data analysis

PLS (Partial least squares regression) is a typical linear algorithm, combining principal component analysis and canonical correlation analysis (Geladi and Kowalski, 1986).

To measure the goodness of the calibration equation under different pre-treatments (Wavelength range, Derivative, Normalization and Standardization). Three indicators were collected including the standard error in cross validation (SECV), the coefficient of determination in calibration (R) and the residual predictive deviation (RPD), which respectively represent the accuracy, sensitivity and stability of the equation. Taking account of these three parameters comprehensively [initially SECV (Cozzolino and Kwiatkowski, 2004), we can pick out the most suitable pre-treatments for the correction equation measuring different quality trait. MSC (Multiplicative Scatter Correction) is a pre-treatment meaning to reduce the scatter effect by correcting every single spectrum based on the univariate linear regression with the average spectrum, SNV (Standard Normal Variate Transform) is also a pretreatment to correct the single spectrum based on the variance of itself. The two are common methods to reduce the noise. To measure how well the calibration model could predict the traits, we used the residual predictive deviation (RPD). The RPD is defined as the standard deviation (S.D.) of the population's reference values divided by the standard error in cross validation for the NIRS calibrations. If the error for estimating a constituent (SECV) is large compared to the spread of that trait in all samples (S.D.), a relatively small RPD is calculated, thereby demonstrating that the NIR calibration model is not robust. In contrast, relatively high RPD values indicate models having greater power to predict the chemical composition. Generally, an RPD greater than three could be considered dependable for prediction purposes.

RESULTS AND DISCUSSION

Testing results of quality traits for samples by traditional methods

The quality traits of the samples used for building the equations are tested by traditional methods (Kjeldahl method for protein content, Soxhlet extractor method for oil content, gas chromatography (GC) for each kinds of oleic acid (Keller, 1961). The result is shown on the Table 1.

The collection of near infrared spectroscopy

Each spectroscopy is scanned and recorded with three repeats, the original spectrum is shown on Fig 1. Along all

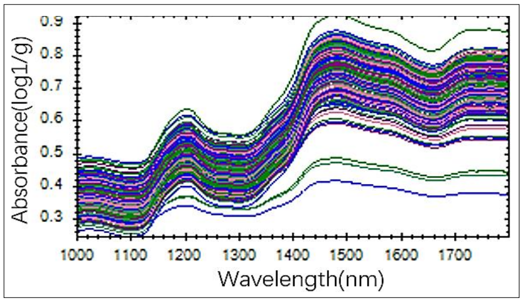

Fig 1: The original scanned infrared spectrum of soybean seeds.

Table 1: Descriptive statistics of the samples used for the development of calibrations for the prediction of five traits.

Constituent	Mean	Standard deviation	Minimum	Maximum
oil content (%)	17.8	1.35	14.6	14.6
protein content (%)	36.4	2.87	27.8	41.7
linoleic acid content (%)	48.2	5.81	32.1	60.6
oleic acid content (%)	33.5	5.15	22.7	46.5
Stearic acid content (%)	2.4	0.28	1.8	2.9

Table 2: The effect of different combination of pre-treatments.

Traits	Wavelength range	Baseline correction	Signal correction	Standardization method	R	SECV	RPDV
Oil content	1000-1799				0.75601	0.907	1.4583
	1000-1799			Mean centering	0.87752	0.697	1.7967
	1000-1799		SNV		0.90399	0.671	1.6708
	1000-1799		SNV	Mean centering	0.89382	0.642	1.7835
	1000-1799		MSC		0.90404	0.669	1.6742
	1000-1799		MSC	Mean centering	0.90404	0.666	1.6742
	1000-1799	First derivative			0.88952	0.696	1.662
	1000-1799	First derivative		Mean centering	0.8979	0.633	1.9282
	1000-1799	First derivative	SNV		0.89946	0.631	1.9341
	1000-1799	First derivative	SNV	Mean centering	0.89963	0.629	1.9381
	1000-1799	First derivative	MSC		0.89957	0.63	1.9376
	1000-1799	First derivative	MSC	Mean centering	0.89957	0.629	1.9376
	1140-1760				0.85828	0.736	2.2271
	1140-1760			Mean centering	0.89325	0.638	1.718
	1140-1760		SNV		0.88627	0.654	1.7861
	1140-1760		SNV	Mean centering	0.88437	0.659	1.713
	1140-1760		MSC		0.88764	0.651	1.7756
	1140-1760		MSC	Mean centering	0.88813	0.65	1.7775
	1140-1760	First derivative			0.87882	0.73	2.1467
	1140-1760	First derivative		Mean centering	0.90114	0.641	1.9256
	1140-1760	First derivative	SNV		0.90729	0.632	1.9855
	1140-1760	First derivative	SNV	Mean centering	0.90721	0.631	1.981
	1140-1760	First derivative	MSC		0.90729	0.632	1.9814
	1140-1760	First derivative	MSC	Mean centering	0.90709	0.631	1.9824
Protein content	1000-1799	0	0	0	0.91877	0.917	1.8749
	1000-1799			Mean centering	0.87844	0.998	1.6453
	1000-1799		SNV		0.9376	0.854	2.4553
	1000-1799		SNV	Mean centering	0.93864	0.838	2.4199
	1000-1799		MSC		0.93753	0.855	2.4558
	1000-1799		MSC	Mean centering	0.93753	0.85	2.4558
	1000-1799	First derivative			0.92583	0.915	2.2061
	1000-1799	First derivative		Mean centering	0.93461	0.818	2.2915
	1000-1799	First derivative	SNV		0.93439	0.833	2.1997
	1000-1799	First derivative	SNV	Mean centering	0.93445	0.83	2.1961
	1000-1799	First derivative	MSC		0.934	0.836	2.1967
	1000-1799	First derivative	MSC	Mean centering	0.934	0.832	2.1967
	1140-1760				0.90839	0.939	1.875
	1140-1760			Mean centering	0.9117	0.901	2.2124
	1140-1760		SNV		0.93596	0.839	2.5182
	1140-1760		SNV	Mean centering	0.93577	0.811	2.5085
	1140-1760		MSC		0.934	0.836	2.1967
	1140-1760		MSC	Mean centering	0.934	0.832	2.1967
	1140-1760	First derivative			0.92815	0.881	1.909
	1140-1760	First derivative		Mean centering	0.93956	0.813	2.3015
	1140-1760	First derivative	SNV		0.93184	0.824	2.2084
	1140-1760	First derivative	SNV	Mean centering	0.93181	0.826	2.2123
	1140-1760	First derivative	MSC		0.93159	0.825	2.2191
	1140-1760	First derivative	MSC	Mean centering	0.93112	0.827	2.1962

Table 2: Continue...

Table 2: Continue...

linoleic acid	1000-1799	0	0	0	0.93121	2.369	2.0034
	1000-1799			Mean centering	0.93773	2.241	2.0777
	1000-1799	First derivative			0.94733	2.449	2.1237
	1000-1799	First derivative		Mean centering	0.95545	2.31	2.1416
	1000-1799	First derivative	SNV		0.95168	2.394	2.1381
	1000-1799	First derivative	SNV	Mean centering	0.95174	2.389	2.1348
	1000-1799	First derivative	MSC		0.95174	2.389	2.1348
	1000-1799	First derivative	MSC	Mean centering	0.95171	2.388	2.1347
	1000-1799		SNV		0.92947	2.46	2.0163
	1000-1799		SNV	Mean centering	0.94237	2.269	2.111
	1000-1799		MSC		0.93913	2.367	2.0846
	1000-1799		MSC	Mean centering	0.93913	2.362	2.0846
	1140-1760				0.91903	2.517	1.9033
	1140-1760			Mean centering	0.93018	2.344	1.9982
	1140-1760		SNV		0.93129	2.349	2.0669
	1140-1760		SNV	Mean centering	0.92816	2.379	2.0291
	1140-1760		MSC		0.93207	2.33	2.0749
	1140-1760		MSC	Mean centering	0.93153	2.329	2.0816
	1140-1760	First derivative			0.9379	2.512	1.9849
	1140-1760	First derivative		Mean centering	0.95198	2.304	2.1575
	1140-1760	First derivative	MSC		0.92798	2.472	1.9966
	1140-1760	First derivative	MSC	Mean centering	0.91834	2.52	2.0418
	1140-1760	First derivative	SNV		0.91897	2.513	2.0436
	1140-1760	First derivative	SNV	Mean centering	0.92767	2.478	1.9901
Oleic acid	1000-1799	0	0	0	0.93945	1.913	2.1686
	1000-1799			Mean centering	0.94306	1.851	2.1399
	1000-1799	First derivative			0.94808	1.874	2.2108
	1000-1799	First derivative		Mean centering	0.9499	1.789	2.2666
	1000-1799	First derivative	SNV		0.92664	2.04	2.0534
	1000-1799	First derivative	SNV	Mean centering	0.92658	2.04	2.0557
	1000-1799	First derivative	MSC		0.92658	2.041	2.0541
	1000-1799	First derivative	MSC	Mean centering	0.92658	2.04	2.0541
	1000-1799		MSC		0.93786	1.968	2.1086
	1000-1799		MSC	Mean centering	0.93786	1.968	2.1086
	1000-1799		SNV		0.93783	1.969	2.1086
	1000-1799		SNV	Mean centering	0.94658	1.849	2.2265
	1140-1760				0.93718	1.928	2.0466
	1140-1760			Mean centering	0.93767	1.902	2.0035
	1140-1760		SNV		0.93942	1.908	2.1414
	1140-1760		SNV	Mean centering	0.93868	1.941	2.1238
	1140-1760		MSC	Mean centering	0.93971	1.906	2.1591
	1140-1760		MSC		0.93995	1.903	2.1519
	1140-1760	First derivative			0.93937	1.918	2.321
	1140-1760	First derivative		Mean centering	0.94118	1.881	2.3145
	1140-1760	First derivative	SNV		0.94224	1.869	2.2877
	1140-1760	First derivative	SNV	Mean centering	0.94277	1.86	2.2985
	1140-1760	First derivative	MSC		0.92256	2.055	2.0726
	1140-1760	First derivative	MSC	Mean centering	0.94254	1.862	2.3051
Stearic acid	1000-1799	0	0	0	0.70784	0.201	1.0963
	1000-1799			Mean centering	0.72494	0.195	1.0881
	1000-1799		SNV		0.83011	0.17	1.0798

Table 2: Continue...

Table 2: Continue...

1000-1799		SNV	Mean centering	0.84029	0.17	1.0626
1000-1799		MSC		0.83006	0.17	1.08
1000-1799		MSC	Mean centering	0.83006	0.169	1.08
1000-1799	First derivative			0.74012	0.192	1.073
1000-1799	First derivative		Mean centering	0.67296	0.21	1.0936
1000-1799	First derivative	SNV		0.78559	0.181	1.0677
1000-1799	First derivative	SNV	Mean centering	0.78677	0.181	1.0665
1000-1799	First derivative	MSC		0.78572	0.181	1.0672
1000-1799	First derivative	MSC	Mean centering	0.78573	0.181	1.0672
1140-1760				0.87903	0.167	1.1825
1140-1760			Mean centering	0.81035	0.175	1.0344
1140-1760		SNV		0.82403	0.174	1.055
1140-1760		SNV	Mean centering	0.82419	0.174	1.0517
1140-1760		MSC		0.82223	0.175	1.0516
1140-1760		MSC	Mean centering	0.80793	0.177	1.0295
1140-1760	First derivative			0.84619	0.17	1.1331
1140-1760	First derivative		Mean centering	0.84204	0.168	1.1534
1140-1760	First derivative	SNV		0.87947	0.168	1.1847
1140-1760	First derivative	SNV	Mean centering	0.87924	0.168	1.1805
1140-1760	First derivative	MSC		0.87913	0.168	1.1829
1140-1760	First derivative	MSC	Mean centering	0.87903	0.167	1.1825

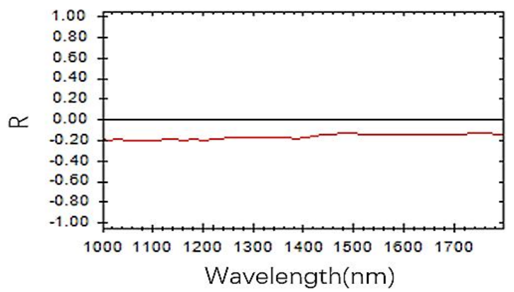

Fig 2: The range of R of the original spectrum on linoleic acid content.

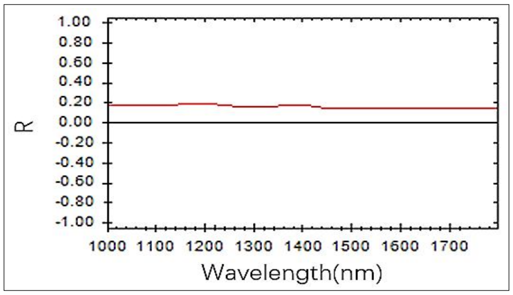

Fig 3: The range of R of the original spectrum on oleic acid content.

the samples, we chose 102 of them is used to train and 10 of them is used to prove. Each sample is scanned repeatedly for 3 times. Every repeat is paired with the value measured by traditional methods to reduce error. As a result, we have 306 training data pairs and 30 proving data pairs.

The building of calibration equation and the effect of different pre-treatment

The original spectrum of samples contains the information of the quality traits, however, it also takes a lot of distractive information caused by scattering and other factors. Thus, we need pre-treatments to adjust the spectrum to enhance the right signal and ignore the wrong signal, but not all these treatments have positive effect on different quality traits, to measure different quality traits, we may need different pre-treatments to convey the information. One pre-treatment might be efficient to enhance the information of oil content, on the other hand, it may weaken the signal of protein content. In one word, pre-treatments are some kind of math methods adjusting the spectrum to excavate the information from it as far as possible. And the result are as follows shown in (Table 2). The different effect of those pre-treatments is shown on the (Fig 1-5), we can see that in the same range of spectrum for different quality traits, the R appears to be different too. So we may say that different range of the spectrum surly carry different information. And for the pre-treatments, we can see huge effects on how the treatments effect the equation in (Table 2), probably the pre-treatments that cause the best parameters might also be the best combination to clear the signal. All the combinations of pre-treatments shown in (Table 2) contains the pre-treatment of Savitzky-Golay smoothing which is not mentioned on the

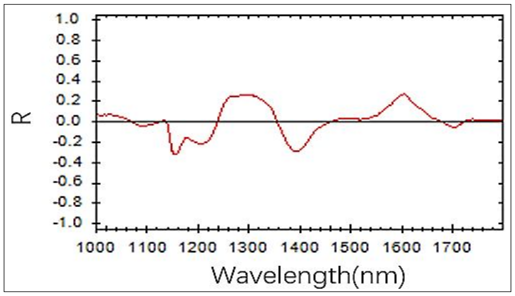

Fig 4: The effect on R of the most suitable pre-treatments on linoleic acid content.

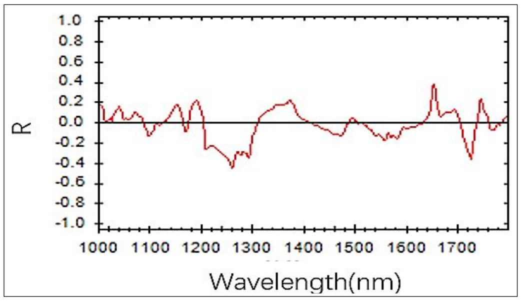

Fig 5: The effect on R of the most suitable pre-treatments on oleic acid content.

table. In the table, four variables are set to find out the best pre-treatments combination for each quality trait.

CONCLUSION

Mainly considered the parameter of SECV, we find that the best combination of pre-treatments for oil content is the combination of first derivative, MSC and mean centering in the range of 1000-1799 nm. For protein content it is the combination of SNV and Mean centering in the range of 1140-1760 nm. For oleic acid content, it is only Mean centering in the range of 1000-1799 nm. For linoleic acid content, it is the combination of First derivative and Mean centering in the range of 1000-1799 nm. For stearic acid content, it is the original spectrum in the range of 1140-1760 that have the best parameters. Those calibration equations and the rough discussions may have some positive effect for the development on both soybean breeding and chemometrics. Besides, we can also see some interesting clue around the (Fig 2-5), for the R range of oleic acid content and linoleic acid content, we can see totally opposite curves from the two. Even after the pre-treatments, R of the same range still appears to be so. As a result, we have enough reasons to conclude that there is a competitive relation between linoleic acid content and oleic acid content.The difference of the curve might be the cause of the difference of structures between unsaturated bond and saturated bond.

In addition, we can see most of the best pre-treatment combination contains SNV but not MSC, consider the fact that SNV correct the spectrum based on the variance while MSC correct the spectrum based on the univariate linear regression, which means that SNV correct the single spectrum only considering the variance within one single spectrum, but the MSC correct the single spectrum considering the relation between all the spectrums and the average spectrum. So we can say that if the noise is predictable and ordered, MSC have an advantage, if not, SNV is better.

Giving the parameters in this study, we may say that the noise in the process of scanning might not be regular.

Author contributions

Conceptualization, Jian Zhang; methodology, Ronghan Ma and Weiran Gao; Software, Weiran Gao; Validation, Aohua Jiang; Formal analysis Jiaqi Liu; Resources, Jian Zhang; Data curation, Pingting Tan and Fang Liu; Writing-original draft preparation, Weiran Gao; Writing-review and editing, Weiran Gao and Jian Zhang; Visualization, Jian Zhang and Weiran Gao; Supervision, Jian Zhang; Project administration, Jian Zhang and funding acquisition, Jian Zhang. All authors have read and agreed to the published version of the manuscript.

Funding

This study was supported by Chongqing Technology Innovation and Application Development Special Key Project (cstc2021jscx-gksbX0011); Collection, Utilization and Innovation of Germplasm Resources by Research Institutes and Enterprises of Chongqing (cqnyncw-kqlhtxm) and Southwest University experimental technology research project (syj2023005).

Institutional review board statement

Not applicable.

Data availability statement

The data presented in this study are available from the first author upon request.

Conflict of interest

The authors declare no conflict of interest.

REFERENCES

Clemente, T.E., Cahoon, E.B. (2009). Soybean oil: Genetic approaches for modification of functionality and total content. Plant Physiology. 151(3): 1030-1040. DOI: 10.2307/40537933.

Cortes, V., Blasco, J., Aleixos, N., Cubero, S., Talens, P. (2019). Monitoring strategies for quality control of agricultural products using visible and near-infrared spectroscopy: A review. Trends Food Sci. Technol. 85: 138-148. DOI: 10.1016/j.tifs.2019.01.015.

Cozzolino, D., Kwiatkowski, M., Parker, M., Cynkar, W., Dambergs, R., Gishen, M., Herderich, M. (2004). Prediction of phenolic compounds in red wine fermentations by visible and near infrared spectroscopy. Analytica Chimica Acta. 513(1): 73-80. DOI: 10.1016/j.aca.2003.08.066.

Geladi, P., Kowalski, B.R. (1986). Partial least-squares regression: A tutorial. Analytica Chimica Acta. 185: 1-17.

Jung, S., Rickert, D.A., Deak, N.A., Aldin, E.D., Recknor, J., Johnson, L.A., Murphy, P.A. (2003). Comparison of Kjeldahl and Dumas methods for determining protein contents of soybean products. Journal of the American Oil Chemists' Society. 80: 1169-1173. DOI: 10.1007/s11746-003-0837-3.

Keller, R. A. (1961). Gas chromatography. Scientific American. 205(4): 58-67.

Kinney, A.J., Knowlton, S. (1998). Designer Oils: The High Oleic Acid Soybean. In: Genetic Modification in the Food Industry. [Roller, S., Harlander, S. (eds)]. Blackie Academic, London. pp 193-213.

McClure, W. F. (2003). 204 years of near infrared technology: 1800-2003. Journal of Near Infrared Spectroscopy. 11(6): 487-518. DOI: 10.1255/jnirs.399.

Medic, J., Atkinson, C., Hurburgh, C.R.J. (2014). Current knowledge in soybean composition. Journal of the American Oil Chemists Society. 91(3): 363-384.

Nicolaï, B.M., Beullens, K., Bobelyn, E., Peirs, A., Saeys, W., Theron, K.I., Lammertyn, J. (2007). Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review. Postharvest Biology and Technology. 46(2): 99-118. DOI: 10.1016/j.postharvbio.2007.06.02.

Nikolić, A., Jović, B., Csanady, S., Petrović, S. (2007). N-H...O Hydrogen bonding: FT IR, NIR and 1H NMR study of N-Methylpropionamide-Cyclic ether systems. J. Mol. Struct. 834-836, 249-252. DOI: 10.1016/j.molstruc.2006.11.003.

Qiao, L., Mu, Y., Lu, B., Tang, X. (2021). Calibration maintenance application of near-infrared spectrometric model in food analysis. Food Rev. Int. 1-17. DOI: 10.1080/87559129.2021.1935999.

Zambelli, A. (2020). Current status of high oleic seed oils in food processing. Journal of the American Oil Chemists' Society. 98: 129-137. DOI: 10.1002/aocs.12450.